STIC-ILL

From: Goldberg, Jeanine
Sent: Thursday, December 20, 2001 3:05 PM
To: STIC-ILL
Subject: *please pull hgt1 references*

1. AMERICAN JOURNAL OF MEDICAL GENETICS, (1999 Dec 15) 88 (6) 694-9.
   Journal code: 3L4; 7708900. ISSN: 0148-7299.

2. HUMAN MOLECULAR GENETICS, (2000 Jul 22) 9 (12) 1753-8.
   Journal code: BRC. ISSN: 0964-6906.

3. GENE, (2001 May 30) 270 (1-2) 69-76.
   Journal code: FOP; 7706761. ISSN: 0378-1119.

4. Genomics Vol 32 (1) pages 75-85 1996

5. Genomics Vol 25 No. 3, pages 707-715 1995.

6. ARCHIVES OF NEUROLOGY, (2001 Oct) 58 (10) 1649-53.
   Journal code: 80K; 0372436. ISSN: 0003-9942.

THANK YOU

Jeanine Enewold Goldberg
1655
CM1--12D11
Mailbox-- 12E12
306-5817

GENOMICS 32, 75–85 (1996)
ARTICLE NO. 0078

Development of a Screening Set for New $(CAG/CTG)_n$ Dynamic Mutations

JULIE M. GASTIER,* THOMAS BRODY,*,[1] JACQUELINE C. PULIDO,*,†,[1] THOMAS BUSINGA,‡ SARA SUNDEN,‡ XINTONG HU,§ SHANAK MAITRA,§ KENNETH H. BUETOW,¶ JEFFREY C. MURRAY,‖ VAL C. SHEFFIELD,‡ MARK BOGUSKI,** GEOFFREY M. DUYK,*,†,[1] AND THOMAS J. HUDSON§,[2]

*Department of Genetics, †Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115; ‡Department of Pediatrics, University of Iowa, Iowa City, Iowa 52242; §Center for Genome Research, Whitehead Institute/Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; ¶Fox Chase Cancer Research Center, Philadelphia, Pennsylvania 19111; ‖Departments of Pediatrics and Biological Sciences, University of Iowa, Iowa City, Iowa 52245; and **National Center for Biotechnology Information, Bethesda, Maryland 20894

Received September 5, 1995; accepted November 22, 1995

The expansion of a $(CAG/CTG)_n$ triplet repeat has been found to be associated with at least seven genetic diseases, suggesting that this mechanism of disease may be fairly common. To accelerate the discovery of new loci containing $(CAG/CTG)_n$ triplet expansions, we have isolated numerous genomic clones containing this class of repeats. We have developed 338 sequence-tagged sites (STSs) containing $(CAG/CTG)_n$ repeat sequences. Two hundred ninety-nine STSs were unambiguously assigned to chromosomes, and 89 of the total were assigned to YACs. The 141 STSs that were developed based on $(CAG/CTG)_n$ repeats of at least seven units were genotyped on four reference CEPH individuals to estimate their polymorphic quality.

INTRODUCTION

Expansions of trinucleotide repeats have been found to be associated with at least nine human diseases, including Fragile X syndrome, FRAXE mental retardation, myotonic dystrophy, Kennedy syndrome, Huntington's disease, spinocerebellar ataxia type 1, dentatorubral-pallido-luysian atrophy, Haw River disease, and Machado-Joseph disease (Verkerk *et al.*, 1991; La-Spada *et al.*, 1991; Brook *et al.*, 1992; Mahadevan *et al.*, 1992; Fu *et al.*, 1992; The Huntington's Disease Collaborative Research Group, 1993; Orr *et al.*, 1993; Knight *et al.*, 1993; Koide *et al.*, 1994; Nagafuchi *et al.*, 1994; Burke *et al.*, 1994; Kawaguchi *et al.*, 1994). Each of these diseases, as well as two additional fragile sites [FRAXF (Nancarrow *et al.*, 1994) and FRA16A (Parrish *et al.*, 1994)], is associated with the expansion of a $(CGG)_n$ or $(CAG/CTG)_n$ repeat.

In an attempt to facilitate the discovery of other disease states associated with trinucleotide repeat expansions, we have initiated a global approach to cloning trinucleotide repeats from genomic DNA. We have developed STSs based on all 10 classes of trinucleotide repeats generated from marker-enriched small insert genomic libraries. The results of a survey of the 10 classes of trinucleotide repeats for their usefulness as new genetic markers have been presented elsewhere (Gastier *et al.*, 1995). Here, we present the results of our efforts to generate large numbers of STSs based on $(CAG/CTG)_n$ repeats, since this class has been found to be associated with the majority of the diseases listed above (all but the fragile sites). These STSs are a valuable screening set in the search for disease mutations suspected to be caused by a trinucleotide repeat expansion, such as other neurodegenerative diseases and diseases showing evidence of anticipation.

MATERIALS AND METHODS

Hybridization conditions. P1 clones (DuPont-Merck) and marker-enriched small insert clones [prepared as described previously (Pulido and Duyk, 1994)] were picked into 96-well plates and replicated onto MAGNA-Nylon membranes (Micron Separations, Inc.). Cosmid clones (Stratagene) were lifted directly off plates. After being grown and fixed on the membranes, clones were screened using the Quick-Light hybridization system (FMC Corp.). Hybridization and washes were performed at 58°C, and control clones were used on all primary and secondary screenings to enhance the number of positives that were picked that had a repeat length of at least five units.

Development of sequence-tagged sites. Hybridization-positive clones were subjected to single-pass cycle sequencing using the M13 (-21) and/or SP6 dye primer kits (Applied Biosystems, Inc.) with the ABI373 automated sequencer. Template DNA was prepared using the Magic Minipreps kit (Promega Corp.). Duplicate clones were detected using Sequencher (GeneCodes Corp.). Primers flanking the repeat were chosen using the Primer program (MIT/Whitehead Insti-

[1] Present address: Millennium Pharmaceuticals Incorporated, Cambridge, MA 02139.

[2] To whom correspondence should be addressed at Center for Genome Research, Whitehead Institute/MIT, 1 Kendall Square Building 300, Room 525, Cambridge, MA 02139. Telephone: (617) 252-1912; Fax: (617) 252-1902.

0888-7543/96 $18.00

tute) as implemented by the CHLC (Cooperative Human Linkage Center) primer pipeline. All primers were selected to have a T_m close to 60°C. For information on the CHLC pipeline server, send a blank e-mail message to primer-server@chlc.org.

Polymorphism analysis. Primers were genotyped on four reference CEPH individuals (1331 01, 1331 02, 1408 01, and 1408 02) under standard PCR conditions. Two children of 1331 01 and 1331 02 (1331 03 and 1331 04) were also typed to ensure a Mendelian inheritance of the alleles.

Chromosome assignment and YAC screening. Tentative localization of each STS to a specific human chromosome was accomplished by PCR-based screening of the National Institute of General Medical Sciences somatic cell hybrid mapping panels 1 and 2. These STSs are being included in the YAC mapping effort at the Whitehead Institute/MIT Center for Genome Research (Bell *et al.,* 1995). The STS content data are freely available via the web server at http://www.genome.wi.mit.edu.

BLAST searches. Various "masking" procedures were applied to the query sequences prior to database homology searching. First, the $(CAG/CTG)_n$ sites themselves (along with other regions of low compositional complexity such as homopolymeric tracts) were identified and masked using the NSEG program (Wootton and Federhen, in press). Masking consists of replacing individual nucleotides with the character "X," which is ignored by the BLAST family of programs (Schuler *et al.,* 1995). These modified query sequences were then screened against a database of vector sequences using the BLASTN program with the E parameter set to 1e-06. All regions of the query sequence matching vector sequences were then masked using the xblast utility (Claverie and States, 1993). These doubly masked query sequences were then searched against the nonredundant nucleotide sequence database using BLASTN with E at 1e-06 as before. The minimum BLAST score that was accepted for homology was 1e-37. Query sequences containing *Alu* repetitive elements were identified by matches to "ALU WARNING" entries in GenBank (Schuler *et al.,* 1995) and by analyzing the alignment outputs. All BLAST searches were carried out using the BLAST network server and databases at NCBI (Schuler *et al.,* 1995).

Electronic access to data. CHLC maps and marker information are available though several electronic information sources: anonymous ftp (ftp.chlc.org), Gopher (gopher.chlc.org), and World Wide Web (WWW) (http://www.chlc.org). Table 2 is available at the CHLC WWW site or though the Whitehead/MIT Center for Genome Research WWW site (http://www-genome.wi.mit.edu). All sequences have been submitted to GDB and Genbank.

RESULTS

Generation of $(CAG/CTG)_n$-Based STSs

We estimate that there are approximately 1500 loci in the human genome with at least five perfect $(CAG/CTG)_n$ units, based on screening a human genomic cosmid library. In addition, we have screened 800 random human genomic P1 clones to confirm this number. In this screen, 4% of the clones were positive. Assuming a 90-kb insert size for the P1 clones and 3000 Mb as the size of the genome, we calculated that there are approximately 1300 $(CAG/CTG)_n$-containing loci in the genome.

TABLE 1

Recovery of $(CAG/CTG)_n$ STSs

	Number of clones
Primers designed	375 (33.0%)
No suitable primers	24 (2.1%)
Misplaced repeat	209 (18.4%)
Repeat length <5	125 (11.0%)
Poor sequence	67 (5.9%)
Duplicates	338 (29.7%)
Total sequenced	1138 (100%)

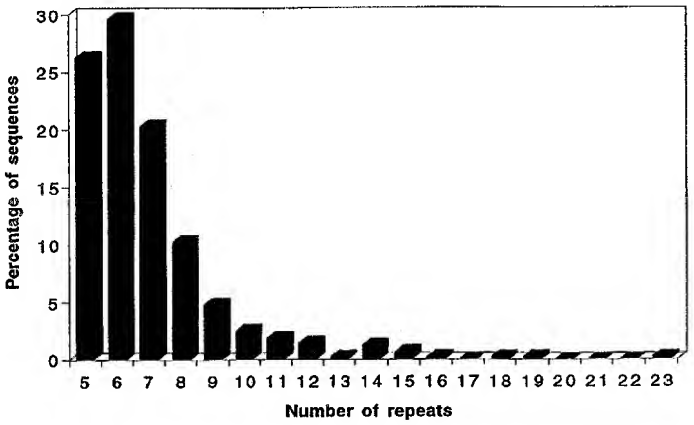

FIG. 1. Distribution of $(CAG/CTG)_n$ repeat lengths in 479 unique clones.

Since the $(CAG/CTG)_n$ class of short tandem repeats is infrequent, on average every 2000–2500 kb, we have utilized small insert-marker-enriched libraries for screening. Hybridization conditions were optimized to yield repeat lengths of 5 or greater based on the discovery that the most common allele of the myotonic dystrophy locus has a repeat length of 5 (Imbert *et al.,* 1993). We have screened a total of approximately 34,000 marker-enriched colonies and are attempting to sequence 2400 positive clones so that primers flanking the repeats can be designed. This paper describes the set of STSs that was developed based on the first 1138 positive clones. The losses encountered in this process are shown in Table 1. It is interesting that this class of repeats is rarely associated with *Alu* repetitive elements, since it is believed that microsatellites may have derived from *Alu* elements (Beckmann and Weber, 1992). Of the 24 failures due to no suitable primers, 11 were due to an *Alu* element flanking the repeat. This means that only 1% of the original clones failed to have primers designed due to an *Alu,* as opposed to other classes of trinucleotide repeats, such as $(ATA)_n$ and $(AAC)_n$, in which 30–40% of the clones failed due to an *Alu* element (Gastier *et al.,* 1995).

Polymorphic Character of the $(CAG/CTG)_n$ STSs

We were able to identify a $(CAG/CTG)_n$ repeat in 479 unique sequences obtained in this survey: the distribution of the length of the repeat in these clones is shown in Fig. 1. The majority of the repeats that were sequenced were short in length. In a previous survey of the 10 classes of trinucleotide repeats, we showed that only repeats of eight or more units tended to be polymorphic (Gastier *et al.,* 1995). This suggested to us

TABLE 2
STSs Developed for (CAG/CTG)$_n$ Repeats

CHLC name (a)	Repeat Sequenced	Alleles 4 CEPHs (b)	YAC data (c)	BLAST homology
Chromosome 1				
GCT1C9	15	1	*	none
GCT1E07	7	1	*	L17906,T02989-STS UT1854, cDNA clone FB19G3
GCT3G04	6			none
GCT3H01	7	1	*	M31776-brain natriuretic protein
GCT4A10	5	1	*	L33798, M87487-calcium channel
GCT4B11	9	1	*	none
GCT4B12	7	1		none
GCT6B08	6			none
GCT6C09	6			none
GCT6H02	7	1		T96519, T84438 cDNA clones
GCT7D01	6			none
GCT7E06	7	1		none
GCT8C07	8	1	*	none
GCT10B12	10	2		none
GCT10C08	6			none
GCT10E12	8	1		none
GCT10G08	11	2		none
GCT11C08	6			none
GCT11G06	6	1		L17904 - Human STS UT1852
GCT11H06	5			none
GCT12H03	5			none
GCT13A06	7	1		none
GCT14E08	5			none
GCT14G04	6			none
GCT15A01	6		*	none
GCT15E11	10	1	*	none
GCT15G02	7	1	*	none
GCT15G08	8	1	*	none
GCT15H08	5			none
GCT16C03	5		*	none
GCT16E08	5		*	none
GCT17A04	6		*	none
GCT17E12	11	?	*	none
Chromosome 2				
GCT1B4	8	1	*	none
GCT3A11	5			none
GCT3C12	5			none
GCT3D12	5			none
GCT4A02	5			none
GCT4A03	5			none
GCT4D02	5			none
GCT5A08	7	1		none
GCT5A09	6			none
GCT5C07	7	2		none
GCT5C11	6			none
GCT5E09	6			none
GCT6D03	6			none
GCT6H03	7	1		none
GCT7C03	7	1		none
GCT8B09 (D2S1397)	10	3		none
GCT9C02	7	1		none
GCT10B07	5			L17898-STS UT1838
GCT10D07	5			none
GCT10F01	7	1		none
GCT10F03	10	1		Z76570-simple DNA seq. wg1a5
GCT11B12	7	2		none
GCT11G10	9	1	*	X76582-simple DNA seq. wg1e10
GCT15B08	6			none
GCT15D07	7	1	*	none
GCT15F06	6			none
GCT15G07	6			none
GCT16H01	7	1	*	none
Chromosome 3				
GCT1A10	15	4		L10376, T07007-CTG-B33 mRNA, EST HFBEC27
GCT1B01	8	1		none
GCT1B06	6	1		none
GCT1D06	6			none
GCT1D8	9	1	*	none
GCT2A09	6			none
GCT2C10	5	1		none
GCT3B12	7	2		none
GCT3C11 (D3S2399)	5		*	none
GCT4B10 (D3S2400)	16	2	*	none
GCT5E11 (D3S2401)	23	7	*	none

TABLE 2—*Continued*

CHLC name (a)	Repeat Sequenced	Alleles 4 CEPHs (b)	YAC data (c)	BLAST homology
GCT6G12	6			none
GCT8B03	7	2		none
GCT8C05	11	2	*	none
GCT8E11	7	2		none
GCT10A08	6			none
GCT10D09	6			none
GCT12G01	9	1		none
GCT14C07	7	1	*	M87731-human simple repeat polymorphism
GCT14D10	7	1		T28895, M97287-EST 59645, SATB1 mRNA
GCT15G11	6			none
GCT16A08	6			none
GCT17B01	9	1	*	none
GCT17B07	7	2	*	none
Chromosome 4				
GCT1B3	7	1	*	none
GCT2C08	5	1		none
GCT4B02	6			none
GCT6B12	5			none
GCT6F01	5			none
GCT6F03 (D4S2430)	12	4	*	none
GCT7C02	6			none
GCT7G03	7	2		none
GCT8D06 (D17S1292)	10	1	*	L17909-STS UT1861
GCT10B09	5	1		none
GCT12H11	5			none
GCT13F01	6			none
GCT14E02	7	2	*	none
GCT15D08	8	1	*	none
GCT16B04	9	3	*	none
GCT16C02	5		*	none
GCT17D01	5		*	none
Chromosome 5				
GCT1A01	6			none
GCT5A12	6			none
GCT5E05 (D5S1472)	8	2	*	none
GCT6A09	8	1	*	none
GCT6E12	6			none
GCT7E01	5			none
GCT7F10	5			none
GCT10A04	9	1		none
GCT10E06	7	1		none
GCT10F04	10	2	*	none
GCT10G10	8	1	*	none
GCT11E05	7	1		none
GCT11G01	6			none
GCT11H05	6			none
GCT15A05	6			none
GCT15B06	5			none
GCT15E04	5			none
Chromosome 6				
GCT4A11	5			none
GCT4B05 (D6S1014)	11	4	*	none
GCT5A01	6			none
GCT5A02	8	1	*	none
GCT5E07 (D6S1015)	11	2	*	none
GCT6G02 (D6S1058)	8	3	*	none
GCT8G05	5			none
GCT10C05	6			none
GCT11E01	5			none
GCT12B05	8	1		L18099-STS UT2507, X73969-wg1f4 repeat region
GCT12B12	5			none
GCT12G04	5			none
GCT16B08	8	2	*	none
GCT16D06	7	2	*	none
GCT16F02	8	1	*	none
Chromosome 7				
GCT9C05	7	2		none
GCT10E08	13	?		none
GCT10F06	5			none
GCT13H07	7	2	*	none
GCT14A05	9	1	*	L17905-STS UT1853
GCT14B10	5			none
GCT15G01	6			none
GCT16H03	5			none

TABLE 2—*Continued*

CHLC name (a)	Repeat Sequenced	Alleles 4 CEPHs (b)	YAC data (c)	BLAST homology
Chromosome 8				
GCT4E02	6			none
GCT5C04	6			none
GCT5H01	5			none
GCT6G07	7	1		none
GCT7F01	5			none
GCT9A01	5			none
GCT10D12	7	1		none
GCT10E01	8	1		none
GCT10F09	8	1	*	L17749-STS UT1022
GCT10H04	5			none
GCT13F07	9	3		none
GCT15A02	6			none
GCT17F04	5		*	none
Chromosome 9				
GCT3G05	9	0	*	none
GCT8A01	7	1		none
GCT8B04	5			none
GCT8G09	6			none
GCT11F04	8	1		none
GCT13B03	6			none
GCT14H05	6			none
GCT16D08	6		*	none
GCT16E06	14	1	*	none
GCT16G03	8	1	*	none
GCT17B09	11	3	*	none
GCT17C06	5		*	none
Chromosome 10				
ACT3E01	11	2	*	none
GCT1C06	6			none
GCT3A04	8	1		none
GCT3E03	6			none
GCT3F05	5			none
GCT8C03	8	1		none
GCT8C11	5			none
Chromosome 11				
GCT5G08	6			none
GCT7G08	5			none
GCT8E07	8	1		none
GCT10A01	5			none
GCT13C12	6	1		Z15459-partial cDNA clone 20B07
GCT13D04	5			none
GCT14C12	6	2		X14972, X53773-mouse, rat alpha-adaptin
GCT14E11	6			none
GCT16A03	5			none
GCT16B07	5		*	none
GCT16B10	5		*	none
GCT16F07	6		*	none
GCT16G07	6		*	none
GCT17D11	8	1	*	none
Chromosome 12				
GCT1A11	6	2		none
GCT1C5	8	1	*	none
GCT5A05	5	1		J04182-lamp-1mRNA
GCT5D05	5			none
GCT6E07 (D12S1072)	12	3	*	none
GCT8B07	9	1	*	none
GCT8G12	9			none
GCT9C01	7	1		none
GCT12G10	6			none
Chromosome 13				
ACT3F12	6	5		none
GCT1C11	5			none
GCT4G05	6			none
GCT7B03	19	?	*	R18580-cDNA clone 30262
GCT7B05	7	1		none
GCT7F02	6			none
GCT13E04	5			none
GCT16A11	6			none
GCT16C05	6		*	none
GCT16F03	6		*	none
GCT17F01	5		*	none

TABLE 2—*Continued*

CHLC name (a)	Repeat Sequenced	Alleles 4 CEPHs (b)	YAC data (c)	BLAST homology
Chromosome 14				
GCT2B12	6			none
GCT2C07	5			none
GCT6H01	7	1	*	none
GCT7H01	7	1		none
GCT8B05	7	1		none
GCT8D03	7	1		none
GCT9B10	6			none
GCT11B02	5			none
GCT13E12	6			none
GCT14E06	8	1		none
GCT15G09	5			none
GCT16B06	6		*	none
Chromosome 15				
GCT1C8	8	2	*	none
GCT2C03	6			none
GCT3B06	6			none
GCT4G01	8	1	*	none
GCT5C12	5			none
GCT6B06	11	1		X76569-human simple DNA seq clone wg1a4
GCT6F04	6			none
GCT7C09	9	2	*	L17911-STS UT1869
GCT10E11	5			none
GCT11A04	8	1		none
GCT11A05	5			none
GCT12B11	7	1		none
GCT13E05	5			none
GCT13E07	6			none
GCT13F05	5	1		L35568, S70721-islet 2, islet 1 mRNA (many species)
GCT14H07	7	3	*	none
Chromosome 16				
GCT2C05	8	1		none
GCT3B03	5			none
GCT3B05	8	1	*	none
GCT3B11	8	2		none
GCT7F04	5			none
GCT7F11	6			none
GCT10B02	8	1		none
GCT10D03	7	1		none
GCT10E09	6			none
GCT13F06	5			none
GCT13F09	7	1	*	none
GCT14B11	7	1	*	none
GCT15A12	6			none
GCT15C04	8	1		none
GCT15D10	9	1	*	none
GCT16F05	6		*	none
GCT16F08	6		*	L26339-human autoantigen mRNA
Chromosome 17				
GCT1E1 (D17S1291)	6		*	none
GCT6E11	10	2	*	none
GCT7A04	5			none
GCT7D11	6			R82424, R31127 cDNA clones
GCT10C02	6	1		none
GCT10D04	13	2		D29801-mouse mRNA ORF
GCT13F02	5			none
GCT14B05	5			none
GCT15E02	6			none
GCT16D12	8	1		none
GCT17C04	7	1		none
GCT17F07	6			none
Chromosome 18				
GCT3A09	6	1		none
GCT3E06 (D18S880)	7	3		none
GCT3G01	8	1		none
GCT5D07 (D18S852)	8	2		none
GCT6F12	6	1		none
GCT6G01	5			none
GCT7G01	6			none
GCT13D05	6			none
Chromosome 19				
GCT2C12	6	1		none
GCT4A09	7	1		none
GCT5G03	5			none
GCT13A07	8	1		none

TABLE 2—*Continued*

CHLC name (a)	Repeat Sequenced	Alleles 4 CEPHs (b)	YAC data (c)	BLAST homology
GCT15A10	9	?		none
Chromosome 20				
GCT10C10	14?	5		none
GCT10F11	9	1		none
GCT11G03	8	2		none
GCT11G09	7	1		none
GCT12F08	7	1		none
GCT13B02	6			none
GCT13C07	8	1		none
GCT14B04	6			none
GCT14G11	5			none
Chromosome 21				
GCT2B10	6			none
GCT12D02	6			none
GCT15A04	6	1		M34876-amyloid-beta gene (APP)
Chromosome 22				
GCT6F02	7	1		none
GCT16D01	7	1		none
X Chromosome				
GCT4C10	7	2		none
GCT5D10	5			none
GCT6D06	6			none
GCT7D06	7	1		none
GCT13B01	5			none
GCT13D12	6			none
GCT14E12	5			none
No Chromosome-High Background				
GCT1C07	6			none
GCT1D04	5	0		X52611-AP-2 transcription factor
GCT1D12	6	0		none
GCT1E10	8	0		none
GCT2B02	5			none
GCT2C04	7			none
GCT3A10	5			X81699-B. taurus sodium dependentphosphate transporter
GCT3B04	5			none
GCT3C10	6			T25829, R73200 cDNA clones CAG-isl 6, 156123
GCT3D10	6			none
GCT3E07	7	0		L17984-STS UT2163
GCT3H07	7	0		M96859-dipeptidyl aminopeptidase like protein mRNA
GCT4B04	5			none
GCT4B08	6	0		none
GCT5A10	5			none
GCT5D01	5			none
GCT5G04	6			none
GCT8D08	6			none
GCT10A11	6			none
GCT10G11	12	3		none
GCT10H03	5			none
GCT10H06	18	1		none
GCT12B01	6			none
GCT12B04	6			none
GCT12D06	5			none
GCT12G03	6			none
GCT13A10	5			none
GCT13C01	4			none
GCT13G09	5			none
GCT14A01	11	0		none
GCT14B02	8	1		none
GCT14C05	5			none
GCT14F06	12	?		none
GCT14H06	5			none
GCT15B10	6			none
GCT15D03	5			none
GCT16D03	5			M28432, M21772-human keratin type 16, keratin pseudogene
GCT16G06	6			none
GCT16G12	6			none

Note. STSs are arranged by chromosome, as assigned by somatic cell hybrid panel mapping.

[a] Name of clone and primers listed in CHLC database. ACT clones were identified in a screen for that class of repeats, but also contain a (CAG/CTG)$_n$ repeat.

[b] Number of alleles in four CEPH individuals (?, results could not be interpreted due to smears or >2 alleles/individual).

[c] (*) Indicates that the STS has been mapped to a YAC. The data is available through the World Wide Web at http://www-genome.wi.mit.edu.

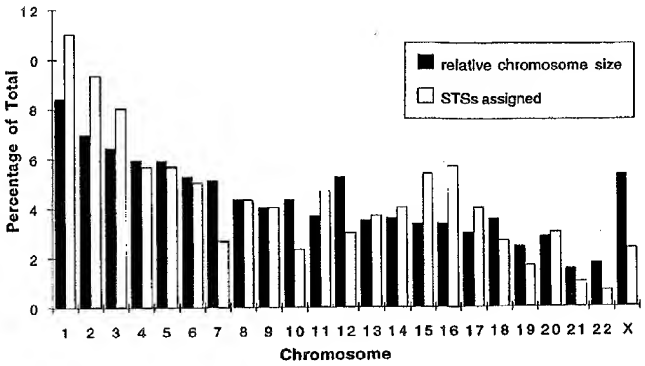

FIG. 2. Distribution of chromosomal assignments for 299 (CAG/CTG)$_n$-containing STSs. The genetic length of a chromosome was divided by the total genetic length of the genome, based on the most recent CHLC maps (Murray *et al.*, 1994).

that most of the (CAG/CTG)$_n$ STSs would not be highly polymorphic. To test this, all STSs based on a repeat length of 7 or greater were tested for their informativeness in four reference CEPH individuals, assuming that an STS with three alleles in the four individuals will tend to be polymorphic in the general population. This is the same method employed to estimate the polymorphic quality of a given STS based on an (AC)$_n$ repeat (Weissenbach *et al.*, 1992). This method is only a crude estimate of informativeness, but has been useful for determining the STSs that are most likely to be polymorphic. Additional individuals must be typed to determine true heterozygosity frequencies. STSs based on repeats of lengths 5 and 6 were generally not typed on the CEPH individuals for cost and efficiency reasons. A small number of these shorter repeats were tested, but most were monoallelic, as shown in Table 2. However, the STSs are included in the (CAG/CTG)$_n$ survey because of the data on myotonic dystrophy, as stated above.

STSs Developed

A summary of the STSs developed in this survey is shown in Table 2. A full table including primer sequences is available at the CHLC or Whitehead/MIT World Wide Web site (see Materials and Methods). The chromosome assignments were obtained by somatic cell hybrid mapping for 299 (88%) of the STSs. Figure 2 shows the chromosome distribution of the assigned STSs vs a normalized length for each chromosome. Thirty-nine (12%) STSs could not be assigned to a specific chromosome due to high mouse or hamster background. We have included these STSs at the end of Table 2 because they may have primers designed in conserved regions, suggestive of coding sequence. The sequence of each clone is available through www.chlc.org and may allow the selection of alternate primers to amplify these loci.

Other Loci Identified in the Screen

Table 3 shows a list of additional loci that were detected in the screen and identified by BLAST searches. These include loci for which primers were developed but did not amplify; sequences where the repeat length was less than 5; clones that were positive by hybridization, but the sequence did not extend into the repeat region; and sequences where primer design was impossible due to lack of flanking sequence.

DISCUSSION

We have developed a set of STSs to be used for screening diseases suspected to be caused by the expansion of a (CAG/CTG)$_n$ repeat. Since seven diseases have been found to be associated with a trinucleotide repeat of this class, it is likely that other diseases have a similar mechanism. By utilizing marker-enriched libraries, we have generated 2400 clones that are positive for this type of repeat. This survey includes the results of analyzing the first 1138 of these clones, and we estimate that 800 of these were unique from each other. These and future STSs can be obtained through the World Wide Web at www.chlc.org.

Frequency estimations indicate that this class of repeats is 50-fold less frequent than the (CA)$_n$ repeats We have found that the repeats tend to be shorter in length and less polymorphic than those in other trinucleotide repeat classes (Gastier *et al.*, 1995). This suggests that there may be an evolutionary restriction on the mutation rate of this class.

The (CAG/CTG)$_n$-containing STSs described in this paper complement the two previous approaches that have been used to identify trinucleotide repeats that may cause disease. Several groups have screened the databases and cDNA libraries for genes containing trinucleotide repeats (Riggins *et al.*, 1992; Li *et al.*, 1993). Based on these efforts, the gene responsible for DRPLA and Haw River syndrome has been cloned (Koide *et al.*, 1994; Nagafuchi *et al.*, 1994; Burke *et al.*, 1994), demonstrating the utility of a random screening approach for identifying loci associated with trinucleotide repeats. In this paper, the sequencing of genomic DNA instead of cDNAs allowed for primer design without the concern for intron/exon boundaries. In addition, we may have identified some triplet-containing gene sequences that are not expressed at levels high enough to be detected in cDNA libraries. Since (CAG/CTG)$_n$ repeats have been shown to be relatively rare in introns and concentrated in coding sequence (Stallings, 1994), it is likely that many of the STSs that we have identified are located in exon sequences.

We have identified several loci that were also detected by the groups searching the database and cDNA libraries for (CAG/CTG)$_n$ repeats (Riggins *et al.*, 1992; Li *et al.*, 1993). These include brain natriuretic protein (GCT3H01), transcription factor AP-2 (GCT1D04), and CTG-B33 (GCT1A10). We have also detected a putative homologous locus of the Machado-Joseph protein (GCT4A06). This clone shares 82% nucleic acid identity near the repeat region of MJD1, but has a different repeat motif [(GCCGCT)8 (GCT)3]. It is likely to be the

TABLE 3

Other Loci Isolated in the Screen

CHLC name	Repeat sequenced	BLAST homology
GCT1A06	(AGC)7	T83552-cDNA clone 111124
GCT1B12	None	M26434-HPRT gene
GCT2A02	(CAG)3--(CAG)3	U03495-transcription factor LSF-ID mRNA
GCT3A05	None	R28340, M94046-cDNA clone 134756, MAZ mRNA
GCT3E01	(AGC)5	R82424-cDNA clone 149143
GCT3F03	(CAG)3--(CAG)2--(CAG)2	M37760, X637550 high-sulfur keratin (many species)
GCT3H02	(CTG)5	K00534, K01903-c-myc
GCT3H05	(GCT)3 CGG (GCT)3	T27013, U17280-cDNA clone LLAB132A10, StAR mRNA
GCT4A06	(GCCGCT)8 (GCT)3	MJD1 protein (homologous locus)
GCT4C04	(GCA)4	ribosomal nontranscribed spacer
GCT5F03	(CAG)4	X14720-c-fms proto-oncogene for CSF-1 receptor
GCT5G07	(CAG)8	S62539-insulin-receptor substrate 1
GCT6A08	None	R59748-cDNA clone 42349, zeta protein
GCT6B03	(GCT)7	R06288-cDNA clone 126292
GCT6D02	(CAG)4	M16801-mineralocorticoid receptor mRNA
GCT7A02	(GCT)6	T84379-cDNA clone 111196
GCT7D09	(CTG)4	X54134-HPTP epsilon mRNA
GCT7E08	None	S47244-HB2B-high sulfur keratin B2B
GCT7G07	(AGC)4	Z26491-catechol o-methyltransferase
GCT8A09	(CAG)6	T67179-cDNA clone 66628
GCT8D02	(GCT)7	U03398-receptor 4-1BB ligand mRNA
GCT8G04	None	M23492-leukocyte common antigen T200 (CD45, LCA)
GCT8H06	(CAG)6--(CAG)7	U23862-mcag32 chromosome 7 CTG repeat region
GCT8H08	None	F11952, M86700-cDNA clone c-33g12, phospholipase A2 mRNA
GCT9D02	None	T25372-cDNA clone BL29-2
GCT10A03	(CTG)6	T08157, T34263-ESTs 06048, 65013
GCT10D10	(GCT)6	R06288-cDNA clone 126292
GCT11E01	(GCT)5	R57209-cDNA clone F1503
GCT12A08	(GCT)5	I08101, I08711, M2446-patents, SFTP3
GCT12A10	(AGC)4	U00115-zinc-finger protein bcl-3
GCT12G09	(GCT)7	R39715-cDNA clone 136883
GCT13C12	(TGC)6	Z15459, T39585-cDNAs, 20B07, 60900
GCT13F08	None	J05272-IMP dehydrogenase type 1 mRNA
GCT14A10	None	M91585-peregrin mRNA
GCT15B09	(GCT)9	X82209-MN1 mRNA
GCT15C09	(AGC)3--(AGC)4	T27046-cDNA clone LLAB212E08
GCT15H11	(CTG)4	U25765-chromosome 17q21 mRNA clone
GCT16A10	(GCT)4	X52560-nuclear factor NF-IL6
GCT16E12	(GCT)5	X15357-mRNA for natriuretic peptide receptor (ANP-A)
GCT16H04	(CAG)4	L17913-STS UT1873
GCT17F03	(TGC)7	M33782-TFEB protein mRNA
GCT17F06	(GCT)5	M78249-EST 00397

Note. All clones were positive by hybridization. The list includes clones for which the primers did not amplify, a repeat was not reached by sequencing, the repeat was shorter than five perfect units, or primer design was impossible due to lack of flanking sequence.

same as one of the MJD-like sequences described in the initial MJD discovery (Kawaguchi *et al.,* 1994). We do not know whether this locus is expressed, but we have determined that the locus tentatively maps to chromosome 8 by somatic cell hybrid mapping (unpublished data).

The repeat expansion detection (RED) assay (Schalling *et al.,* 1993) is another technique that has been used to facilitate cloning of other expanded loci. RED, which allows the identification of a triplet expansion in a given individual, requires no prior knowledge of flanking sequence. RED has allowed the detection of novel expansions of (CTG)$_n$, (ATG)$_n$, (CCT)$_n$, (CTT)$_n$, and (TGG)$_n$ trinucleotide repeats in the genomic DNA of normal individuals (Schalling *et al.,* 1993; Linblad *et al.,* 1994) as well as the presence of (CAG/CTG)$_n$ triplet expansions in the genomic DNA of individuals with bipolar affective disorder and schizophrenia (Linblad *et al.,* in press; O'Donovan *et al.,* 1995). Targeted cloning of long CTG triplet molecules detected by RED has proven elusive so far, in part due to the difficulty in propagating long trinucleotide repeat sequences in bacterial and yeast host systems. STSs generated in this screen are candidate markers to test in individuals identified by RED to have putative (CAG/CTG)$_n$ expansions.

These STSs will be useful in the continuing search for disease mutations caused by a trinucleotide repeat expansion. In addition, in cells with DNA repair defects, genes that contain (GCT)/CTG)$_n$ and other microsatellite repeats may lead to disease without large

expansions. For example, it has been shown that, in cells with a repair defect, the gene encoding the RII subunit of the TGF-β receptor has accumulated mutations in a polyadenine tract that may inactivate the gene (Markowitz *et al.,* 1995). This suggests that all loci containing microsatellite repeats are candidates for disease-causing agents in some cancers, and the STSs described here may help to identify some of these as well.

ACKNOWLEDGMENTS

We thank the following people for their help on this project: Jelveh Ghazizadeh, George Church, Gary Gryan, and Keith Robison (Harvard Medical School); Kerry Wiles, Dee Even, Molly Wise, and Cynthia Wichtman (Iowa); and John Wooten for advice on NSEG parameters. This work was supported in part by NIH Grant P50HG 00835. Thomas J. Hudson is a recipient of a Clinician-Scientist award from the Medical Research Council of Canada.

REFERENCES

Beckmann, J. S., and Weber, J. L. (1992). Survey of human and rat microsatellites. *Genomics* 12: 627–631.

Bell, C. J., Nieuwenhuijsen, B. W., Barnoski, B., Budarf, M. L., Buetow, K. H., Campbell, K., Colbert, A., Collins, J., Desjardins, P. R., DeZwaan, T., Eckman, B., Fischbeck, K., Foote, S., Hart, K., Hiester, K., Van Het Hoog, M. J., Hopper, E., McDermid, H. E., Overton, C., Reeve, M. P., Searls, D. B., Watson, E., Winston, R., Valmiki, V. H., Nussbaum, R. L., Lander, E. S., Emanuel, B. S., and Hudson, T. J. (1995). Integration of physical, breakpoint and genetic maps of chromosome 22. Localization of 587 yeast artificial chromosomes with 238 mapped markers. *Hum. Mol. Genet.* 4: 59–69.

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J.-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J., and Housman, D. E. (1992). Molecular basis of myotonic dystropy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* 68: 799–808.

Burke, J. R., Wingfield, M. S., Lewis, K. E., Roses, A. D., Lee, J. E., Hulette, C., Pericak-Vance, M. A., and Vance, J. M. (1994). The Haw River Syndrome: Dentatorubropallidoluysian atrophy (DRPLA) in an African-American family. *Nature Genet.* 7: 521–524.

Claverie, J.-M., and States, D. J. (1993). Information enhancement methods for large scale sequence analysis. *Comput. Chem.* 17: 191–201.

Fu, Y.-H., Pizzuti, A., Fenwick, R. G., Jr., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., De Jong, P., Wieringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F., and Caskey, C. T. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* 255: 1256–1258.

Gastier, J. M., Pulido, J. C., Sunden, S., Brody, T., Buetow, K. H., Murray, J. C., Weber, J. L., Hudson, T. J., Sheffield, V. C., and Duyk, G. M. (1995). Survey of trinucleotide repeats in the human genome: Assessment of their utility as genetic markers. *Hum. Mol. Genet.* 4: 1829–1836.

Imbert, G., Kretz, C., Johnson, K., and Mandel, J.-L. (1993). Origin of the expansion mutation in myotonic dystrophy. *Nature Genet.* 4: 72–76.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S., and Kakizuka, A. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nature Genet.* 8: 221–227.

Knight, S. J. L., Flannery, A. V., Hirst, M. C., Campbell, L., Christodoulou, Z., Phelps, S. R., Pointon, J., Middleton-Price, H. R., Barnicoat, A., Pembrey, M. E., Holland, J., Oostra, B. A., Bobrow, M., and Davies, K. E. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell* 74: 127–134.

Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F., and Tsuiji, S. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nature Genet.* 6: 9–13.

La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E., and Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 352: 77–79.

Li, S.-H., McInnis, M. G., Margolis, R. L., Antonarakis, S. E., and Ross, C. A. (1993). Novel triplet repeat containing genes in human brain: Cloning, expression, and length polymorphisms. *Genomics* 16: 572–579.

Linblad, K., Nylander, P.-O., De Bruyn, A., Sourey, D., Zander, C., Engstrom, C., Holmgren, G., Hudson, T., Chotai, J., Mendlewicz, J., Van Broeckhoven, C., Schalling, M., and Adolfsson, R. (1995). Detection of expanded CAG repeats in Bipolar Affective Disorder using the repeat expansion detection (RED) method. *Neurobiol. Disease* 2: 55–62.

Linblad, K., Zander, C., Schalling, M., and Hudson, T. (1994). Growing triplet repeats. *Nature Genet.* 7: 124.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-MacDonald, J., De Jong, P. J., Wieringa, B., and Korneluk, R. G. (1992). Myotonic Dystrophy mutation: An unstable CTG repeat in the 3′ untranslated region of the gene. *Science* 255: 1253–1255.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R. S., Zborowska, E., Kinzler, K. W., Vogelstein, B., Brattain, M., and Willson, J. K. V. (1995). Inactivation of the Type II TGF-β receptor in colon cancer cells with microsatellite instability. *Science* 268: 1336–1338.

Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V. C., Sunden, S., Duyk, G., Weissenbach, J., Gyapay, G., Dib, C., Morissette, J., Lathrop, G. M., Vignal, A., White, R., Matsunami, N., Gerken, S., Melis, R., Albertsen, H., Plaetke, R., Odelberg, S., Ward, D., Dausset, J., Cohen, D., and Cann, H. (1994). A comprehensive human linkage map with centimorgan density. *Science* 265: 2049–2054.

Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama N., Tanaka, Y., Kikushima, H., Umino, K., Kurosawa, H., Furukawa, T., Nihei, K., Inoue, T., Sano, A., Komure, O., Takahashi, M., Yoshizawa, T., Kanazawa, I., and Yamada, M. (1994). Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* 6: 14–18.

Nancarrow, J. K., Kremer, E., Holman, K., Eyre, H., Doggett, N. A., Le Paslier, D., Callen, D. F., Sutherland, G. R., and Richards, R. I. (1994). Implications of FRA16A structure for the mechanism of chromosomal fragile site genesis. *Science* 264: 1938–1941.

O'Donovan, M. C., Guy, C., Craddock, N., Murphy, K. C., Cardno, A. G., Jones, L. A., Owen, M. J., and McGufin, P. (1995). Expanded CAG repeats in schizophrenia and bipolar disorder. *Nature Genet.* 10: 380–381.

Orr, H. T., Chung, M., Banfi, S., Kwiatkowski, T. J., Jr., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P. W., and Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* 4: 221–226.

Parrish, J. E., Oostra, B. A., Verkerk, A. J. M. H., Richards, C. S., Reynolds, J., Spikes, A. S., Shaffer, L. G., and Nelson, D. L. (1994).

Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. *Nature Genet.* 8: 229–235.

Pulido, J. C., and Duyk, G. M. (1994). *In* "Current Protocols in Human Genetics" (N. C. Dracopoli, J. L. Haines, B. R. Korf, D. T. Moir, C. C. Morton, C. E. Seidman, J. G. Seidman, and D. R. Smith, Eds.), Vol. 1, pp. 2.2.1–2.2.33, Wiley, New York.

Riggins, G. J., Lokey, L. K., Chastain, J. L., Leiner, H. A., Sherman, S. L., Wilkinson, K. D., and Warren, S. T. (1992). Human genes containing polymorphic trinucleotide repeats. *Nature Genet.* 2: 186–191.

Schalling, M., Hudson, T. J., Buetow, K. H., and Housman, D. E. (1993). Direct detection of novel expanded trinucleotide repeats in the human genome. *Nature Genet.* 4: 135–139.

Schuler, G. D., Boguski, M. S., and Gish, W. (1995). Sequence similarity searching using the BLAST family of programs. *In* "Current Protocols in Human Genetics" (N. C. Dracopoli, J. L. Haines, B. R. Korf, D. T. Moir, C. C. Morton, C. E. Seidman, J. G. Seidman, and D. R. Smith, Eds.), Vol. 12, pp. 11.3.1–11.3.37, Wiley, New York.

Stallings, R. L. (1994). Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequences: Implications for human genetic diseases. *Genomics* 21: 116–121.

The Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72: 971–983.

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F., Eussen, B. E., van Ommen, G.-J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A., and Warren, S. T. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in Fragile X syndrome. *Cell* 65: 905–914.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. (1992). A second-generation linkage map of the human genome. *Nature* 359: 794–801.

Wootton, J. C., and Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.,* in press.